ATSC Standard:
A/342 Part 3, MPEG-H System

Doc. A/342-3:2022-03
31 March 2022

Advanced Television Systems Committee
1300 I Street, N.W., Suite 400E
Washington, D.C. 20005
202-872-9160

The Advanced Television Systems Committee, Inc. is an international, non-profit organization developing voluntary standards and recommended practices for broadcast television and multimedia data distribution. ATSC member organizations represent the broadcast, professional equipment, motion picture, consumer electronics, computer, cable, satellite, and semiconductor industries. ATSC also develops implementation strategies and supports educational activities on ATSC standards. ATSC was formed in 1983 by the member organizations of the Joint Committee on Inter-society Coordination (JCIC): the Consumer Technology Association (CTA), the Institute of Electrical and Electronics Engineers (IEEE), the National Association of Broadcasters (NAB), the Internet & Television Association (NCTA), and the Society of Motion Picture and Television Engineers (SMPTE). For more information visit www.atsc.org.

Note: The user's attention is called to the possibility that compliance with this standard may require use of an invention covered by patent rights. By publication of this standard, no position is taken with respect to the validity of this claim or of any patent rights in connection therewith. One or more patent holders have, however, filed a statement regarding the terms on which such patent holder(s) may be willing to grant a license under these rights to individuals or entities desiring to obtain such a license. Details may be obtained from the ATSC Secretary and the patent holder.

Implementers with feedback, comments, or potential bug reports relating to this document may contact ATSC at https://www.atsc.org/feedback/.

Revision History

Version	Date
Candidate Standard approved	3 May 2016
Updated CS approved	9 November 2016
A/342-3:2017 Standard approved	3 March 2017
Document number inconsistency fixed on title page and running header	7 March 2017
A/342-3:2021 Standard approved	11 March 2021
A/342-3:2022-03 published (references to ATSC documents updated)	31 March 2022

Table of Contents

1. SCOPE 1
   1.1 Introduction and Background 1
   1.2 Organization 1
2. REFERENCES 1
   2.1 Normative References 1
   2.2 Informative References 2
3. DEFINITION OF TERMS 2
   3.1 Compliance Notation 2
   3.2 Treatment of Syntactic Elements 2
      3.2.1 Reserved Elements 3
   3.3 Acronyms and Abbreviations 3
   3.4 Terms 4
4. MPEG-H AUDIO SYSTEM OVERVIEW 4
   4.1 Introduction 4
   4.2 MPEG-H Audio System Features 4
      4.2.1 Metadata Audio Elements 4
      4.2.2 Multi-Stream Delivery 6
      4.2.3 Audio/Video Fragment Alignment and Seamless Configuration Change 7
      4.2.4 Seamless Switching 7
      4.2.5 Loudness and Dynamic Range Control 8
      4.2.6 Interactive Loudness Control 8
5. MPEG-H AUDIO SPECIFICATION 8
   5.1 Audio Encoding 8
   5.2 Bit Stream Encapsulation 8
      5.2.1 MHAS (MPEG-H Audio Stream) Elementary Stream 8
      5.2.2 ISOBMFF Encapsulation 9
   5.3 Audio Loudness and DRC Signaling 11
   5.4 Audio Emergency Information 11

ANNEX A: DECODER GUIDELINES 12
   A.1 MPEG-H Audio Decoder Overview 12
   A.2 Output Signals 12
      A.2.1 Loudspeaker Output 12
      A.2.2 Binaural Output to Headphones 13
   A.3 Decoder Behavior 13
      A.3.1 Tune-in 13
      A.3.2 Configuration Change 14
      A.3.3 Adaptive Bitrate Switching 14
   A.4 User Interface for Interactivity 14
      A.4.1 Audio Scene and User Interactivity Information 14
      A.4.2 MPEG-H Audio Decoder API for User Interface 15
      A.4.3 User Interface on Systems Level 15
   A.5 Loudness Normalization and Dynamic Range Control 16
   A.6 System Sounds and Voice Assistant Sounds 16

Index of Figures and Tables

Figure 4.1 MPEG-H Audio Scene Information example. 6
Figure 4.2 Example of switching and merging multiple incoming streams. 7
Figure A.1.1 MPEG-H Audio Decoder (features, block diagram, signal flow). 12
Figure A.4.1 MPEG-H Audio Decoder Interface for user interactivity. 15
Figure A.4.2 Interface on systems level for user interactivity. 16
Figure A.6.1 Top level block diagram of a receiver using MPEG-H Audio 17

Table A.6.1 PCM Data and Earcon Metadata Use Description 17

ATSC Standard:
A/342 Part 3, MPEG-H System

1. SCOPE

This document standardizes the MPEG-H Audio system for use in the ATSC 3.0 Digital Television System. It describes the characteristics of the MPEG-H Audio system and establishes a set of constraints on MPEG-H Audio [2] [4] for use within ATSC 3.0 broadcast emissions.

1.1 Introduction and Background

The ATSC 3.0 audio system provides an enhanced feature set and improves upon the capabilities of past ATSC audio systems. The system provides listeners with both a personalized and an immersive experience. The ATSC 3.0 audio system establishes a common framework for multiple Next Generation Audio (NGA) systems, both current and future. The MPEG-H Audio system is one of the NGA systems standardized in ATSC 3.0.

1.2 Organization

This document is organized as follows:

- Section 1 – Outlines the scope of this document and provides a general introduction.
- Section 2 – Lists references and applicable documents.
- Section 3 – Provides definitions of terms, acronyms, and abbreviations for this document.
- Section 0 – Gives an overview of the MPEG-H Audio System for ATSC 3.0.
- Section 5 – Specifies MPEG-H Audio for ATSC 3.0.
- Annex A: – Describes Decoder Guidelines for MPEG-H Audio.

2. REFERENCES

All referenced documents are subject to revision. Users of this Standard are cautioned that newer editions might or might not be compatible.

2.1 Normative References

The following documents, in whole or in part, as referenced in this document, contain specific provisions that are to be followed strictly in order to implement provisions of this Standard.

[1] IEEE: "Use of the International Systems of Units (SI): The Modern Metric System," Doc. SI 10, Institute of Electrical and Electronics Engineers, New York, NY.

[2] ISO/IEC: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," Doc. 23008-3:2019, International Standards Organization / International Electrotechnical Commission, Geneva, Switzerland.

[3] ISO/IEC: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," AMENDMENT 1, Doc. 23008-3:2019, International Standards Organization / International Electrotechnical Commission, Geneva Switzerland.

[4] ISO/IEC: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," AMENDMENT 2, Doc. 23008-3:2019, International Standards Organization / International Electrotechnical Commission, Geneva Switzerland.

[5] ISO/IEC: "MPEG audio technologies – Part 4: Dynamic Range Control," Doc. 23003-4:2020, International Standards Organization / International Electrotechnical Commission, Geneva Switzerland.

[6] IETF: BCP 47, "Tags for Identifying Languages," Internet Engineering Task Force, Reston, VA, September 2009.

2.2 Informative References

The following documents contain information that may be helpful in applying this Standard.

[7] ATSC: "Techniques for Establishing and Maintaining Audio Loudness for Digital Television," Doc. A/85:2013, Advanced Television Systems Committee, Washington, D.C., 12 March 2013. See also A/85:2013 Corrigendum No. 1, "SPL," approved 11 February 2021.

[8] ATSC: "ATSC Standard: A/342 Part 1, Audio Common Elements," Doc. A/342-1:2022-03, Advanced Television Systems Committee, Washington, DC, 31 March 2022.

[9] ITU: "Multichannel stereophonic sound system with and without accompanying picture," Recommendation ITU-R BS.775-3, International Telecommunications Union, Geneva, Switzerland.

[10] ITU: "Algorithms to measure audio programme loudness and true-peak audio level," Recommendation ITU-R BS.1770-4, International Telecommunications Union, Geneva, Switzerland, 2015.

[11] ITU: "Requirements for Loudness and True-Peak Indicating Meters," Recommendation ITU-R BS.1771-1, International Telecommunications Union, Geneva, Switzerland.

[12] ITU: "Advanced sound system for programme production," Recommendation ITU-R BS.2051-2, International Telecommunications Union, Geneva, Switzerland.

[13] IANA: language subtag registry, http://www.iana.org/numbers.html under "Language Tags".

3. DEFINITION OF TERMS

With respect to definition of terms, abbreviations, and units, the practice of the Institute of Electrical and Electronics Engineers (IEEE) as outlined in the Institute's published standards [1] shall be used. Where an abbreviation is not covered by IEEE practice or industry practice differs from IEEE practice, the abbreviation in question will be described in Section 3.3 of this document.

3.1 Compliance Notation

This section defines compliance terms for use by this document:

shall – This word indicates specific provisions that are to be followed strictly (no deviation is permitted).

shall not – This phrase indicates specific provisions that are absolutely prohibited.

should – This word indicates that a certain course of action is preferred but not necessarily required.

should not – This phrase means a certain possibility or course of action is undesirable but not prohibited.

3.2 Treatment of Syntactic Elements

This document contains symbolic references to syntactic elements used in the audio, video, and transport coding subsystems. These references are typographically distinguished by the use of a

different font (e.g., restricted), may contain the underscore character (e.g., sequence_end_code), and may consist of character strings that are not English words (e.g., dynrng).

3.2.1 Reserved Elements

One or more reserved bits, symbols, fields, or ranges of values (i.e., elements) may be present in this document. These are used primarily to enable adding new values to a syntactical structure without altering its syntax or causing a problem with backwards compatibility, but they also can be used for other reasons.

The ATSC default value for reserved bits is '1'. There is no default value for other reserved elements. Use of reserved elements except as defined in ATSC Standards or by an industry standards setting body is not permitted. See individual element semantics for mandatory settings and any additional use constraints. As currently reserved elements may be assigned values and meanings in future versions of this Standard, receiving devices built to this version are expected to ignore all values appearing in currently reserved elements to avoid possible future failure to function as intended.

3.3 Acronyms and Abbreviations

The following acronyms and abbreviations are used within this document.

A/V – Audio/Video
AD – Audio Description
API – Application Programing Interface
ATSC – Advanced Television Systems Committee
BCP – Best Current Practice
BL – Baseline Profile
BRIR – Binaural rendering impulse response
DASH – Dynamic Adaptive Streaming over HTTP
DRC – Dynamic Range Control
EAS – Emergency Alert System
GUI – Graphical User Interface
HDMI – High-Definition Multimedia Interface
HOA – Higher Order Ambisonics
HRTF – Head-Related Transfer Function
HTTP – Hypertext Transfer Protocol
IANA – Internet Assigned Numbers Authority
IEC – International Electrotechnical Commission
IEEE – Institute of Electrical and Electronics Engineers
IP – Internet Protocol
IPF – Immediate Play-out Frame
ISO – International Organization for Standardization
ISOBMFF – ISO Base Media File Format
ITU-R – International Telecommunication Union - Radiocommunication Sector
LC – Low Complexity
MAE – Metadata Audio Element
MHAS – MPEG-H Audio Stream

mhm1– ISO Base Media File Format sample entry type for MPEG-H Audio Stream
MPEG-D – Abbreviation for Parts of ISO/IEC 23003
MPEG-H – Abbreviation for Parts of ISO/IEC 23008
NGA – Next Generation Audio
PCM – Pulse-code modulation
RAP – Random Access Point
RF – Radio Frequency
TV – Television
UI – User Interface

3.4 Terms

The following terms are used within this document.

Audio Description – As defined in A/342 Part 1 [8], "Audio Description" (AD) is the insertion of audio narrated descriptions of a television program's key visual elements into natural pauses between the program's dialogue.

Audio Emergency Information – As defined in A/342 Part 1 [8], "Emergency Information" is data to be presented aurally, such as the reading of a text crawl, and is distinct from Emergency Alert System (EAS) data and audio.

reserved – Set aside for future use by a Standard.

Fragment – One part of a fragmented ISOBMFF file.

Scalable/Layered Audio – An audio system that uses hierarchical techniques to: (1) Provide increasing quality of service with improving reception conditions, and/or (2) Provide different levels of service quality as required by different device types or presentation environments.

4. MPEG-H AUDIO SYSTEM OVERVIEW

4.1 Introduction

The MPEG-H Audio system (ISO/IEC 23008-3 [2] [4]) offers methods for coding of channel-based content, coding of object-based content, and coding of scene-based content (using Higher Order Ambisonics (HOA) as a sound-field representation). An MPEG-H Audio encoded program may consist of a flexible combination of any of the audio program elements that are defined in ATSC A/342-1 Section 5.3.2 [8], namely:

- Channels (i.e., signals for specific loudspeaker positions),
- Objects (i.e., signals with position information) and
- Higher Order Ambisonics, HOA (i.e., sound field signals).

4.2 MPEG-H Audio System Features

4.2.1 Metadata Audio Elements

MPEG-H Audio uses a set of static metadata, the "Metadata Audio Elements" (MAE), to define an "Audio Scene". An Audio Scene represents an Audio Program as defined in ATSC A/342-1 Section 4 [8].

Audio Objects are associated with metadata that contains all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts.

The metadata (MAE) is structured in several hierarchy levels. The top-level element of MAE is the "AudioSceneInfo". Sub-structures of the Audio Scene Info contain "Groups", "Switch Groups", and "Presets". Groups represent Audio Program Components; Presets represent Audio Presentations as defined in ATSC A/342-1 Section 4 [8].

4.2.1.1 Groups of Elements

The concept of an element group enables arranging related element signals that are to be treated together as a unit, e.g., for interactivity in common or for simultaneous rendering. A use case for groups of elements is the definition of channel-based recordings as audio elements (e.g., a stereo recording in which the two signals should only be manipulated as a pair). Grouping of elements allows for signaling of stems and sub-mixes by collecting the included element signals into groups that then can be treated as single objects.

4.2.1.2 Switch Groups of Elements

The concept of a switch group describes a grouping of elements that are mutually exclusive with one another. It can be used to ensure that exactly one of the switch group members is enabled at a time. This allows for switching between, e.g., different language tracks, when it is not sensible to simultaneously enable multiple language tracks.

4.2.1.3 Presets

Presets can be used to offer pre-configured combinations of groups and objects for more convenient user selection. Properties of the groups, like default gain or position can be set differently for each preset. It is not necessary to include all groups and objects in a preset.

4.2.1.4 Personalization and Interactive Control

Using the information in MAE, the MPEG-H Audio system offers listeners the ability to interactively control and adjust various elements of an audio scene within limits set by broadcasters (e.g., to adjust the relative level of dialogue only in a range specified within the AudioSceneInfo structure).

4.2.1.5 MAE Audio Scene Example

Figure 4.1 contains an example of MPEG-H Audio Scene Information with five different groups and one switch group. The switch group contains three commentaries from which to choose – two different English commentaries and one foreign language. Additionally, the user may select the "sound effects" object. In this case, the sound effects object is not a single mono source but rather a multi-channel object with pre-rendered content.

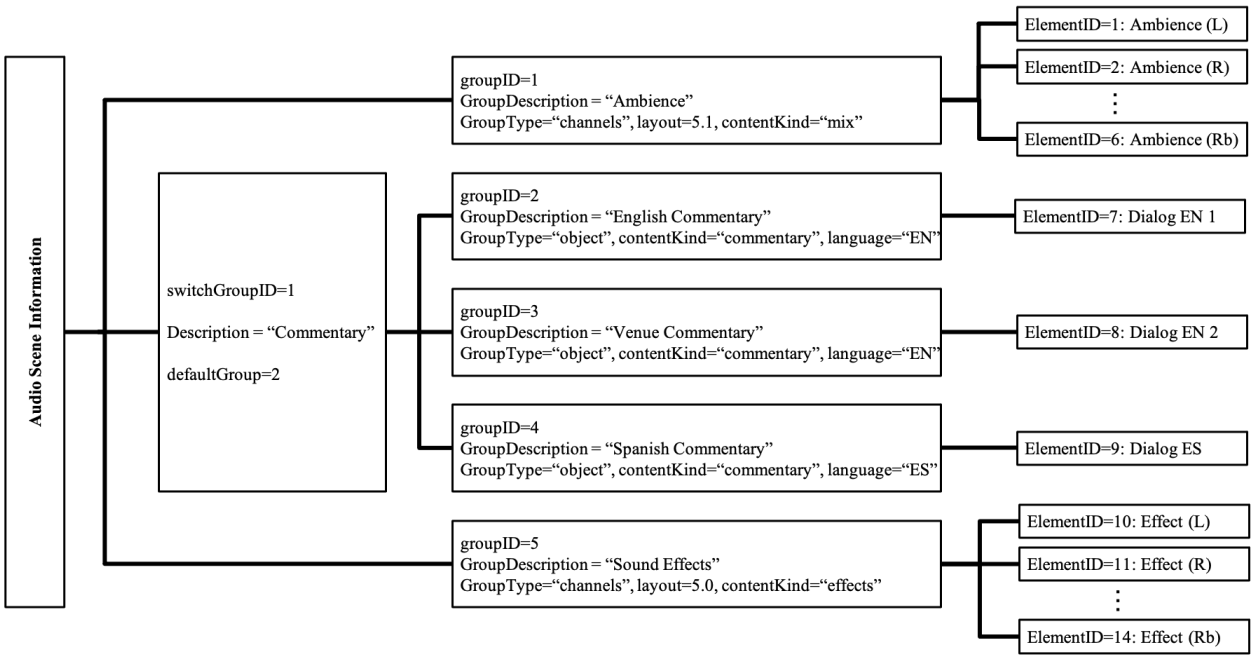

Figure 4.1 MPEG-H Audio Scene Information example.

4.2.2 Multi-Stream Delivery

The multi-stream-enabled MPEG-H Audio system is capable of handling streams delivered in a hybrid environment (e.g., one stream, containing one complete presentation, delivered over Broadcast and one or more additional streams, containing different languages, delivered over Broadband). The MAE information contained in each stream allows the MPEG-H Audio decoder to correctly merge the streams into one stream containing several sub-streams.

Figure 4.2 illustrates an example of such a hybrid delivery scenario, in which, out of several incoming streams, the main stream (stream #0) and the third stream (stream #2) are selected and merged into a single stream, while the second stream (stream #1) is discarded based on information obtained from the systems level.

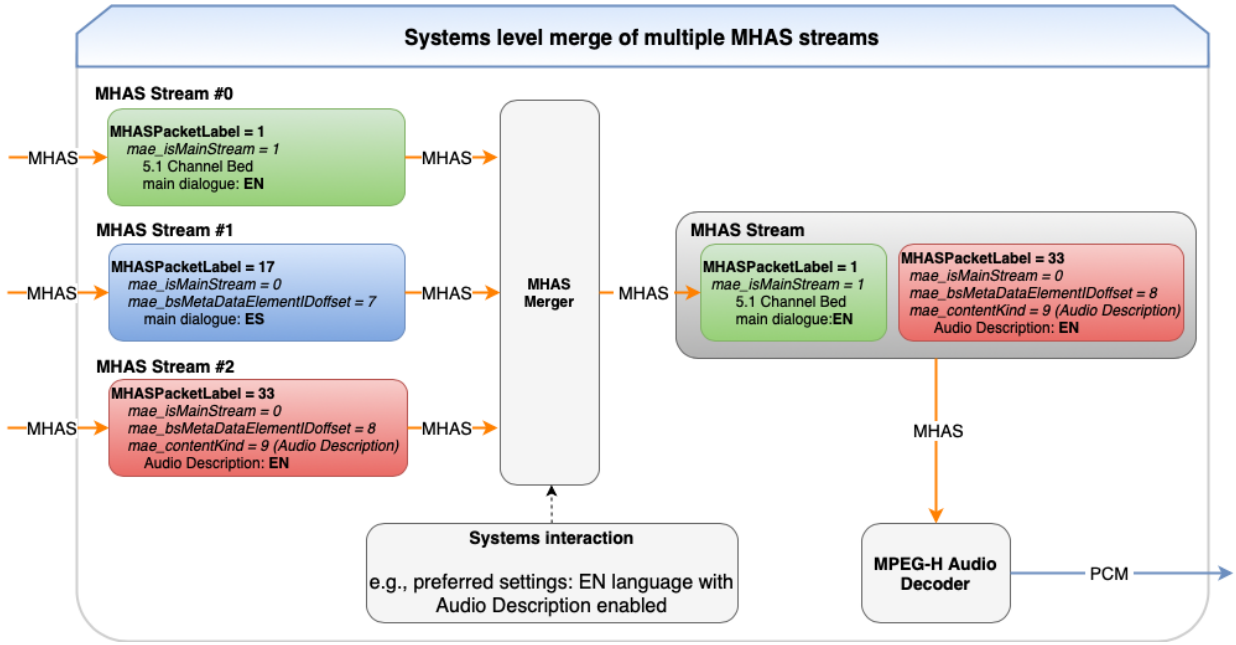

Figure 4.2 Example of switching and merging multiple incoming streams.

4.2.3 Audio/Video Fragment Alignment and Seamless Configuration Change

The MPEG-H Audio system allows for sample-accurate splicing and reconfiguration of the audio stream at Fragment boundaries. For coding efficiency, the audio and video streams usually use different frame rates, however, some applications may require the audio and video streams to be aligned at the end of certain Fragments (e.g., for stream splicing – normally used for commercial insertion). In this case, at least some audio frames usually need to be truncated to a number of audio samples lower than the normal audio frame size.

The MPEG-H Audio system uses a packet structure, called the MPEG-H Audio Stream (MHAS), to transport encoded audio and corresponding metadata. The MHAS packet design guarantees that any MHAS packet payload always is byte-aligned so that the truncation information is easily accessible on the fly and can be easily inserted, removed or modified by, e.g., a stream splicing device.

Another application of the truncation method is changing the configuration of an MPEG-H Audio stream (e.g., a stereo program may be followed by a program with 5.1 channels and additional audio objects). The configuration usually will change at a video frame boundary that is not aligned with the granules of the audio stream. Using the audio sample truncation method, the configuration change can be aligned to the video frame boundary.

4.2.4 Seamless Switching

The MPEG-H Audio system allows for seamless transitions between audio streams by making use of Immediate Play-out Frames (IPFs). Random-access points can be created, for example, by introducing IPFs at arbitrary positions in the audio stream.

IPFs carry additional information of “previous” audio frames and thereby allow seamless transitions between audio streams, e.g., in an adaptive streaming (e.g., DASH) scenario. Having IPFs at segment boundaries allows for glitch-free bitrate adaptation of different audio representations and for on-the-fly reconfiguration of the stream.

4.2.5 Loudness and Dynamic Range Control

The MPEG-H Audio system includes advanced tools for loudness and dynamic range control inherited from MPEG-D DRC. MPEG-D DRC defines a comprehensive and flexible metadata format that includes transmission of loudness metadata according to recommendations ITU-R BS.1770 [7] and ITU-R BS.1771 [11] among others. Further, MPEG-D DRC is compliant to worldwide regulations including those based on ATSC A/85 [7].

The dynamic range control tool offers a rich feature set for accurate adaptation of both complete audio scenes and single elements within audio scenes for different receiving devices and different listening environments. For personalization, the MPEG-H Audio stream can include optional DRC configurations selectable by the user that, e.g., offer improved dialog intelligibility.

In addition to compression provisions, the dynamic range control tool offers broadcaster-controlled ducking of selected audio elements by sending time-varying gain values (e.g., for audio scenes that are combined with voice-overs or Audio Description).

4.2.6 Interactive Loudness Control

MPEG-H Audio enables listeners to interactively control and adjust different elements of an audio scene within limits set by broadcasters. To fulfill applicable broadcast regulations and recommendations with respect to loudness consistency, the MPEG-H Audio system includes a tool for automatic compensation of loudness variations due to user interactions.

5. MPEG-H AUDIO SPECIFICATION

5.1 Audio Encoding

Audio signals shall be encoded into bit streams according to the ISO/IEC 23008-3, MPEG-H Low Complexity (LC) Profile Level 1, 2, or 3, as defined in [2].

If the audio signals are encoded according to the Baseline (BL) Profile restrictions specified in ISO/IEC 23008-3 Amendment 2, Clause 4.8.2.6 [4], the CompatibleProfileLevelSet() config extension element specified in ISO/IEC 23008-3 Amendment 2, Clause 4.8.2.7 [4] shall be present. The mpegh3daProfileLevelIndication and CompatibleSetIndication shall be as set according to Table P.1 in ISO/IEC 23008-3 Amendment 2 [4].

When language is specified, the language tag shall conform to the syntax and semantics defined in IETF BCP 47 [6]. That language tag shall be mapped to the corresponding MAE metadata elements mae_bsDescriptionLanguage and mae_contentLanguage as specified in ISO/IEC 23008-3, Clause 15.3 [2].

Note: The IANA language subtag registry can be found here [13].

5.2 Bit Stream Encapsulation

5.2.1 MHAS (MPEG-H Audio Stream) Elementary Stream

Audio data shall be encapsulated into MPEG-H Audio Stream (MHAS) packets according to ISO/IEC 23008-3, Clause 14 [2].

MHAS packets of all types defined in ISO/IEC 23008-3, Clause 14 [2] and ISO/IEC 23008-3 Amendment 1 [3] may be present in an MHAS elementary stream, except for the following packet types, which shall not be present in the stream:

- PACTYP_CRC16
- PACTYP_CRC32
- PACTYP_GLOBAL_CRC16

- PACTYP_GLOBAL_CRC32

The following packet types may be present in an MHAS elementary stream. If they are present, however, they may be ignored by decoders:

- PACTYP_SYNC
- PACTYP_SYNCGAP

If Audio Scene Information according to ISO/IEC 23008-3, Clause 15 [2] is present, it always shall be encapsulated in an MHAS PACTYP_AUDIOSCENEINFO packet [2]. Audio Scene Information shall not be included in the mpegh3daConfig() structure in the MHAS PACTYP_MPEGH3DACFG packet.

If text labels for Group of Elements, Switch Groups or Presets should be carried within an MPEG-H Audio Stream, they may be encapsulated either as part of the MHAS PACTYP_AUDIOSCENEINFO packet within an mae_Description() structure, or alternatively they may be encapsulated within an MHAS PACTYP_DESCRIPTOR packet carrying an MPEG-H_3dAudio_text_label_descriptor().

If content identifiers should be carried within an MPEG-H Audio Stream, they may be encapsulated in an MHAS PACTYP_MARKER packet with the marker_byte set to "E0".

5.2.2 ISOBMFF Encapsulation

5.2.2.1 MPEG-H Audio Sample Entry

The sample entry "mhm1" shall be used for encapsulation of MHAS packets into ISOBMFF files, according to ISO/IEC 23008-3, Clause 20.6 [2].

The sample entry "mhm2" shall be used in cases of multi-stream or hybrid delivery, i.e., when the MPEG-H Audio Program is split into two or more streams for delivery as described in ISO/IEC 23008-3, Clause 14.6 [2].

If the MHAConfigurationBox() is present, the MPEG-H Profile-Level Indicator mpegh3daProfileLevelIndication in the MHADecoderConfigurationRecord() shall be set to "0x0B", "0x0C", or "0x0D" for MPEG-H Audio Low Complexity Profile Level 1, Level 2, or Level 3, respectively. The Profile-Level Indicator in the MHAS PACTYP_MPEGH3DACFG packet shall be set accordingly.

5.2.2.2 Random Access Point and Stream Access Point

A File Format sample containing a Random Access Point (RAP), i.e., a RAP into an MPEG-H Audio Stream, is a "sync sample" in the ISOBMFF and shall consist of the following MHAS packets, in the following order:

- PACTYP_MPEGH3DACFG
- PACTYP_AUDIOSCENEINFO (if Audio Scene Information is present)
- PACTYP_BUFFERINFO
- PACTYP_MPEGH3DAFRAME

Note that additional MHAS packets may be present between the MHAS packets listed above or after the MHAS packet PACTYP_MPEGH3DAFRAME, with one exception: when present, the PACTYP_AUDIOSCENEINFO packet shall directly follow the PACTYP_MPEGH3DACFG packet, as defined in ISO/IEC 23008-3, Clause 14.4 [2].

Additionally, the following constraints shall apply for sync samples:

- The audio data encapsulated in the MHAS packet PACTYP_MPEGH3DAFRAME shall follow the rules for random access points as defined in ISO/IEC 23008-3, Clause 5.7 [2].
- All rules defined in ISO/IEC 23008-3, Clause 20.6.1 [2] regarding sync samples shall apply.

- The first sample of an ISOBMFF file shall be a RAP. In cases of fragmented ISOBMFF files, the first sample of each Fragment shall be a RAP.
- In case of non-fragmented ISOBMFF files, a RAP shall be signaled by means of the File Format sync sample box "stss", as defined in ISO/IEC 23008-3, Clause 20.2 [2].
- In case of fragmented ISOBMFF files, the sample flags in the Track Run Box ('trun') are used to describe the sync samples. The "sample_is_non_sync_sample" flag shall be set to "0" for a RAP; it shall be set to "1" for all other samples.

5.2.2.3 Configuration Change

A configuration change takes place in an audio stream when the content setup or the Audio Scene Information changes (e.g., when changes occur in the channel layout, the number of objects etc.), and therefore new PACTYP_MPEGH3DACFG and PACTYP_AUDIOSCENEINFO packets are required upon such occurrences. A configuration change usually happens at program boundaries, but it may also occur within a program.

The following constraints apply:

- At each configuration change, the MHASPacketLabel shall be changed to a different value from the MHASPacketLabel in use before the configuration change occurred.
- A configuration change may happen at the beginning of a new ISOBMFF file or Fragment or at any position within the file. In the latter case, the File Format sample that contains a configuration change shall be encoded as a sync sample (RAP) as defined above.
- A sync sample that contains a configuration change and the last sample before such a sync sample may contain a truncation message (PACTYP_AUDIOTRUNCATION) as defined in ISO/IEC 23008-3, Clause 14.4.13 [2].

The usage of truncation messages enables synchronization between video and audio elementary streams at program boundaries. When used, sample-accurate splicing and reconfiguration of the audio stream are possible. If MHAS packets of type PACTYP_AUDIOTRUNCATION are present, they shall be used as described in ISO/IEC 23008-3, Clause 14.4.13 [2].

5.2.2.4 Multi-Stream Delivery

In case of multi-stream delivery (as described in Section 4.2) the Audio Program Components of one Audio Program are not carried within one single MHAS elementary stream, but in two or more MHAS elementary streams.

The following constraints apply for ISOBMFF files or fragments using the sample entry "mhm2":

- The Audio Program Components of one Audio Program are carried in one main MHAS elementary stream, and one or more auxiliary MHAS elementary streams.
- The main MHAS stream shall contain at least the Audio Program Components corresponding to the default Audio Presentation, i.e., the Audio Scene Information is present and exactly one preset shall have the mae_groupPresetID field set to "0", as specified in ISO/IEC 23008-3, Clause 15.3 [2].
- The mae_isMainStream field in the Audio Scene Information shall be set to "1" in the main MHAS stream, as specified in ISO/IEC 23008-3, Clause 15.3 [2]. This field shall be set to "0" in the auxiliary MHAS streams.

- In each auxiliary MHAS stream the mae_bsMetaDataElementIDoffset field in the Audio Scene Information shall be set to the index of the first metadata element in the auxiliary MHAS stream minus one, as specified in ISO/IEC 23008-3, Clause 14.6 and Clause 15.3 [2].
- For the main and the auxiliary MHAS stream(s), the MHASPacketLabel shall be set according to ISO/IEC 23008-3, Clause 14.6 [2].
- The main and the auxiliary MHAS stream(s) that carry Audio Program Components of one Audio Program shall be time aligned.
- In each auxiliary MHAS stream, the random access points (RAP) shall be aligned to the RAPs present in the main MHAS stream.

5.3 Audio Loudness and DRC Signaling

Loudness metadata shall be embedded within the mpegh3daLoudnessInfoSet() structure as defined in ISO/IEC 23008-3, Clause 6.3 [2]. Such loudness metadata shall include at least the loudness of the content rendered to the default rendering layout as indicated by the referenceLayout field (see ISO/IEC 23008-3, Clause 5.3.2 [2]). More precisely, the mpegh3daLoudnessInfoSet() structure shall include at least one loudnessInfo() structure with loudnessInfoType set to "0", whose drcSetId and downmixId fields are set to "0" and which includes at least one methodValue field with methodDefinition set to "1" or "2" (see ISO/IEC 23008-3, Clause 6.3.1 [2] and ISO/IEC 23003-4, Clause 7.3 [5]). The indicated loudness value shall be measured according to local loudness regulations (e.g., ATSC A/85 [7]).

DRC metadata shall be embedded in the mpegh3daUniDrcConfig() and uniDrcGain() structures as defined in ISO/IEC 23008-3, Clause 6.3 [2]. For each included DRC set the drcSetTargetLoudnessPresent field as defined in ISO/IEC 23003-4, Clause 7.3 [5] shall be set to "1". The bsDrcSetTargetLoudnessValueUpper and bsDrcSetTargetLoudnessValueLower fields shall be configured to continuously cover the range of target loudness levels between -31 dB and 0 dB.

Loudness compensation information (mae_LoudnessCompensationData()), as defined in ISO/IEC 23008-3, Clause 15.5 [2], shall be present in the Audio Scene Information if the mae_allowGainInteractivity field (according to ISO/IEC 23008-3, Clause 15.3 [2]) is set to "1" for at least one group of audio elements.

5.4 Audio Emergency Information

The MPEG-H Audio system can insert Audio Emergency Information as an Audio Program Component part of one or more Audio Presentations. If Audio Emergency Information is present than the mae_contentKind field (according to ISO/IEC 23008-3, Clause 15.3 [2]) of the Audio Program Component carrying the Audio Emergency Information shall be set to "12" ("emergency").

Annex A: Decoder Guidelines

A.1 MPEG-H AUDIO DECODER OVERVIEW

MPEG-H Audio offers the possibility to code channel-based content, object-based content and scene-based content, the latter using Higher Order Ambisonics (HOA) for a sound-field representation. Figure A.1.1 gives a brief overview of signal flow in an MPEG-H Audio decoder from bit stream input to loudspeaker or headphone output. As a first step, all transmitted audio signals are decoded by the MPEG-H Audio Core Decoder. Channel-based signals are mapped to the target reproduction loudspeaker layout using the Format Conversion module. Object-based signals are rendered to the target reproduction loudspeaker layout by the Object Renderer. Scene-based content is rendered to the target reproduction loudspeaker layout using associated HOA metadata and an HOA decoder/renderer.

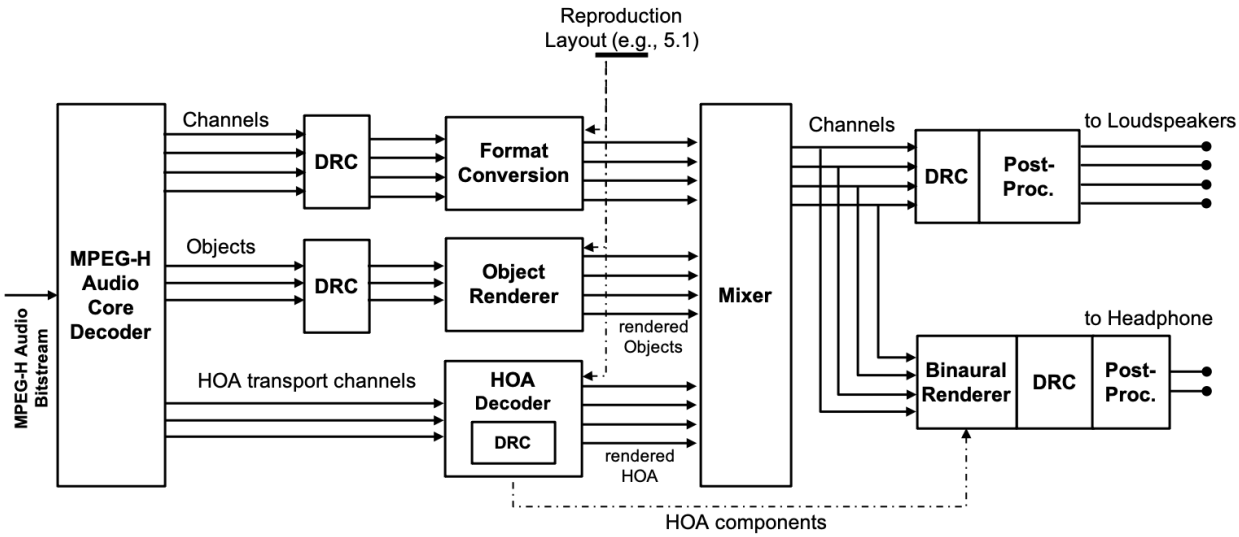

Figure A.1.1 MPEG-H Audio Decoder (features, block diagram, signal flow).

A.2 OUTPUT SIGNALS

A.2.1 Loudspeaker Output

The MPEG-H Audio Decoder is able to render the encoded input signals to loudspeaker output channel signals for any target loudspeaker layout geometry.

The MPEG-H Low Complexity Profile Level 3 specified in ISO/IEC 23008-3 [2] limits the maximum number of output channels to 12.

The receiving device obtains the target loudspeaker layout from external information supplied during system setup and passes that information to the MPEG-H Audio decoder during its initialization, as defined in ISO/IEC 23008-3, Clause 17.2 [2].

Examples for target loudspeaker layouts:

- 2.0 Stereo
- 5.1 Multi-channel audio
- 10.2 Immersive audio

Stereo and 5.1 configurations are defined in ITU-R BS.775-3 [9].

The channel configuration for 10.2 is specified in ITU-R BS.2051 [12].

A.2.2 Binaural Output to Headphones

If the receiving device supports headphone reproduction, the MPEG-H Audio Decoder can render the encoded input signals to a binaural signal for headphone playback, as defined in ISO/IEC 23008-3, Clause 13 [2].

BRIR/HRTF data for binaural rendering can be fed to the decoder as defined in ISO/IEC 23008-3, Clause 17.4 [2].

The value of nBrirPairs in BinauralRendering() may be

- Maximum 2 for level 1,
- Maximum 6 for level 2,
- Maximum 11 for level 3.

The value of kMax in FdBinauralRendererParam() may be equal to or less than 48 (bands). The value of kConv in FdBinauralRendererParam() may be equal to 32. The average of nFilter[k] values in VoffBrirParam() may be less than or equal to 64. The values of rt60[k] in SfrBrirParam() may be less than or equal to 1.0 (sec).

If binaural rendering is activated, measured BRIR positions may be passed to mpegh3daLocalSetupInformation().

All renderer stages thereby are set to a target layout that is equal to the BRIR measurement setup. Preferably, the BRIR measurement positions for standard target layouts 2.0, 5.1, 10.2 and 7.1.4 may be provided.

A.3 DECODER BEHAVIOR

A.3.1 Tune-in

A tune-in happens at a channel change of a receiving device. The audio decoder is able to tune in to a new audio stream at every random access point (RAP).

Starting from the first sync sample (i.e., the RAP) in an ISOBMFF file, the decoder receives ISOBMFF samples containing MHAS packets. As defined above, the sync sample contains the configuration information (PACTYP_MPEGH3DACFG and PACTYP_AUDIOSCENEINFO) that is used to initialize the audio decoder. After initialization, the audio decoder reads encoded audio frames (PACTYP_MPEGH3DAFRAME) and decodes them.

To avoid buffer underflow during reception, the decoder is expected to wait before starting to decode audio frames, as indicated in PACTYP_BUFFERINFO.

Note, that it may be necessary to feed several audio frames into the decoder before the first decoded PCM output buffer is available, as described in ISO/IEC 23008-3, Clause 22 [2].

It is recommended that, on tune-in, the receiving device perform a 100ms fade-in on the first PCM output buffer that it receives from the audio decoder.

A.3.2 Configuration Change

MPEG-H Audio enables seamless configuration change in broadcast and streaming (e.g., DASH) environments.

If a decoder receives an ISOBMFF sample that contains a configuration change, the decoder performs a configuration change according to ISO/IEC 23008-3, Clause 5.5.6 [2].

The configuration change can be detected through the change of the MHASPacketLabel of the packet PACTYP_MPEGH3DACFG compared to the value of the MHASPacketLabel of previous MHAS packets.

Note that a change in the mpegh3daLoudnessInfoSet() and mpegh3daUniDrcConfig() may occur that does not trigger a configuration change.

A.3.3 Adaptive Bitrate Switching

MPEG-H Audio enables seamless bitrate switching in an environment with different representations, i.e., bit streams encoded at different bitrates, of the same content. In the case of a switch between such bit streams, the decoder performs an adaptive switch according to ISO/IEC 23008-3, Clause 5.5.6 [2].

A.4 USER INTERFACE FOR INTERACTIVITY

A.4.1 Audio Scene and User Interactivity Information

If the MPEG-H Audio bit stream enables user interactivity, i.e., it contains an MHAS packet with PACTYP_AUDIOSCENEINFO, the user may change certain aspects of the rendered audio scene during playback, e.g., change the level or position of an audio object.

The Audio Scene Information, as defined in ISO/IEC 23008-3, Clause 15 [2], contains information as to what the user is allowed to change and by how much, e.g., which audio elements are enabled for interactivity and what the maximum allowed changes of those audio elements are (e.g., in terms of gain or position).

The Audio Scene Information also may contain descriptions of "Presets", i.e., combinations of audio elements that are a subset of the complete audio scene, with gains and positions that are different from the default values. Typical examples for such Presets are:

- "Dialogue Enhancement" with increased dialogue signal level and attenuated background signal level, and
- "Live Mix", e.g., for a sports event, with enhanced ambience, an additional ambience object signal, and a muted dialogue object that contains commentary.

The Audio Scene Information also may contain textual labels with descriptions of audio elements or presets that can be used in a Graphical User Interface (GUI).

Changes that result from user interactivity in the GUI are taken into account by the MPEG-H Audio decoder during rendering of the audio scene.

If user interactivity results in gain changes of one or more audio elements in an audio scene, loudness compensation, as defined in ISO/IEC 23008-3, Clause 15.5 [2], is applied.

Two examples are described regarding how Audio Scene Information can be provided to the GUI and how user interactivity information can be provided from the GUI to the audio decoder:

- Through an interface (API) of the MPEG-H Audio decoder, as further described in the following sub-section, or

- Through a separate building block at the systems level of the receiving device, as further described in the subsequent subsection.

A.4.2 MPEG-H Audio Decoder API for User Interface

In the scenario described in this section, an audio decoder has two interfaces: a data interface for the encoded bit stream data and an additional interface for user interactivity information. The additional interface is defined in ISO/IEC 23008-3, Clause 17.7 [2].

As shown in Figure A.4.1, the MPEG-H Audio decoder makes the Audio Scene Information available to an application for usage in a GUI. In return, the MPEG-H Audio decoder receives user interactivity information from the application through the mpegh3daElementInteraction() structure as defined in ISO/IEC 23008-3, Clause 17.7 [2], with the ei_interactionMode field set to "1" only if preset information is present in the bit stream.

Optionally, the MPEG-H Audio decoder may receive an mpegh3daLoudnessDrcInterface() structure as defined in ISO/IEC 23008-3, Clause 14.2.2 [2] for selection of a preferred DRC configuration.

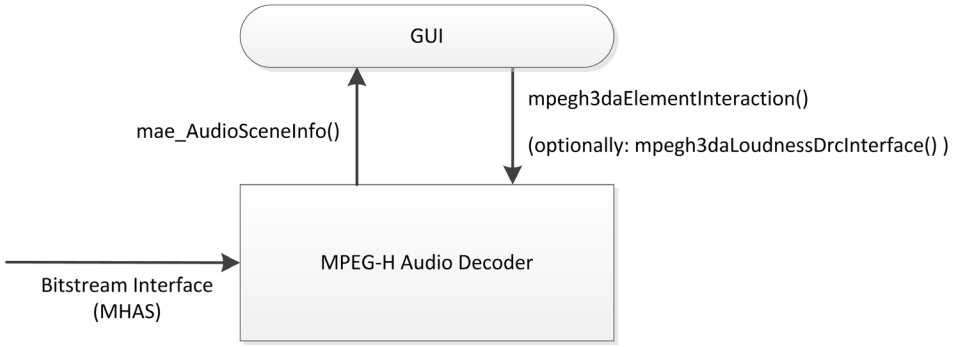

Figure A.4.1 MPEG-H Audio Decoder Interface for user interactivity.

A.4.3 User Interface on Systems Level

In the scenario described in this section, the audio decoder has only the data interface for the encoded bit stream data and no additional interface for user interactivity.

As shown in Figure A.4.2, the Audio Scene Information is extracted from the MPEG-H Audio bit stream at systems level. The "MHAS Extractor" building block parses the MHAS stream, extracts the MHAS PACTYP_AUDIOSCENEINFO packet and makes it available to the application for usage in a GUI.

In return, the "MHAS Embedder" accepts the user interactivity information from the application layer. The user interactivity information is carried in the mpegh3daElementInteraction() structure and is further encapsulated in the MHAS packet of type PACTYP_USERINTERACTION, as defined in ISO/IEC 23008-3, Clause 14.4.9 [2], with the ei_interactionMode field in the mpegh3daElementInteraction() structure set to "1" only if preset information is present in the bit stream.

The "MHAS Embedder" building block embeds the MHAS packet PACTYP_USERINTERACTION into the MHAS packet stream that is fed into the audio decoder. Optionally, the MHAS packet PACTYP_LOUDNESS_DRC, as defined in ISO/IEC 23008-3, Clause 14.4.10 [2], also is embedded into the MHAS packet stream, if the preferred DRC configuration is intended to be changed in the GUI.

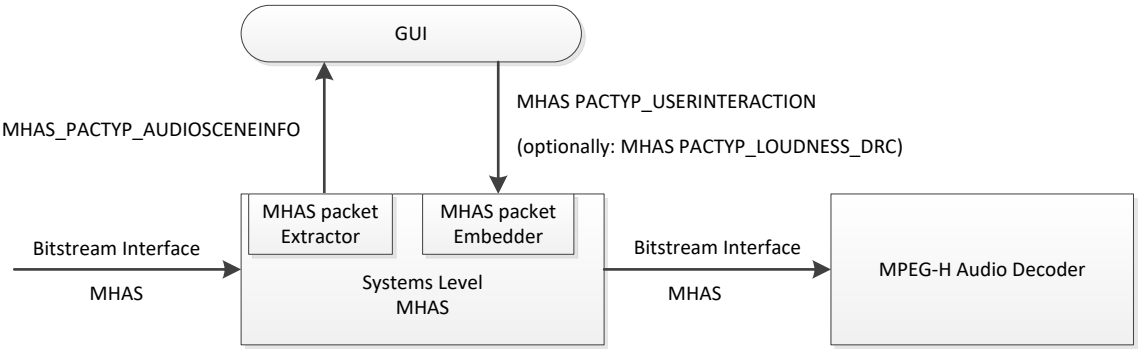

Figure A.4.2 Interface on systems level for user interactivity.

A.5 LOUDNESS NORMALIZATION AND DYNAMIC RANGE CONTROL

The loudness normalization feature of the MPEG-H Audio decoder is expected to be enabled constantly. The target level for normalization typically is fixed and dependent on the specific receiving device type (e.g., A/V receiver, TV set, Mobile device). The maximum target level allowed at the decoder is expected to be restricted to a practical value.

The preferred Loudness and DRC configuration is controlled by the mpegh3daLoudnessDrcInterface() structure, as defined in ISO/IEC 23008-3, Clause 14.2.2 [2] and ISO/IEC 23003-4, Annex B [5]. (Also, see above in the section "User Interface for Interactivity"). The interface is expected to be restricted with respect to changes within the loudnessNormalizationControlInterface() and dynamicRangeControlInterface() structures, in which loudnessNormalizationOn and dynamicRangeControlOn are both expected to have values of "1".

A.6 SYSTEM SOUNDS AND VOICE ASSISTANT SOUNDS

If a receiving device is connected to an external device (e.g., Soundbar), system sounds or voice assistant sounds may be embedded on the fly in the MPEG-H Audio bit stream and delivered to the external device which decodes the MPEG-H Audio bit stream.

The process of embedding the system sounds or voice assistant sounds on the fly in an MPEG-H Audio bit stream is illustrated in Figure A.6.1, where the "PCM to MHAS Embedder" module is responsible to receive the mono or stereo PCM data and create the MHAS packets defined in ISO/IEC 23008-3 Amendment 1, Clause 28 [3], of type:

- PACTYP_EARCON carrying the earconInfo() structure,
- PACTYP_PCMCONFIG carrying the pcmDataConfig() structure and
- PACTYP_PCMDATA carrying the pcmDataPayload() structure.

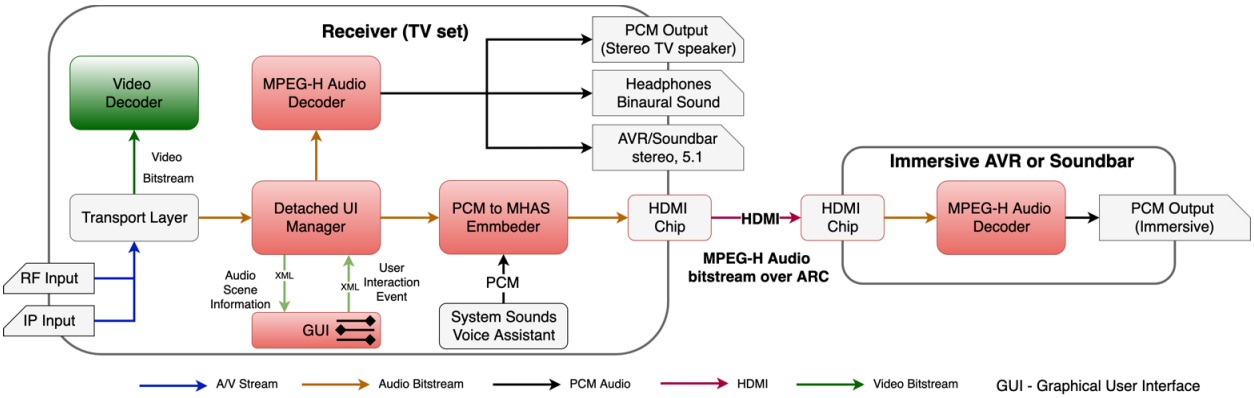

Figure A.6.1 Top level block diagram of a receiver using MPEG-H Audio.

The PCM data and metadata is specified in ISO/IEC 23008-3 Amendment 1, Clause 28 [3] using the constraints listed in Table A.6.1.

Table A.6.1 PCM Data and Earcon Metadata Use Description

Structure	PCM data and Earcon Metadata	Use Description
earconInfo()	bsNumEarcons	Set to "0" for mono PCM sounds, Set to "1" for stereo PCM sounds.
	earconIsIndependent	Set to "1".
	earconType	Set to "5", indicating that data of type PCM is present.
	earconPositionType	Set to "0".
	earcon_CICPspeakerIdx[i]	earcon_CICPspeakerIdx[0] == 2 for mono PCM sounds (i.e., numEarcons == 0); earcon_CICPspeakerIdx[0, 1] == 0, 1 for stereo PCM sounds (i.e., numEarcons == 1);
	earconHasGain	Set to "0".
pcmDataConfig()	bsNumPcmSignals	Same value as bsNumEarcons.
	pcmAlignAudioFlag	Set to "0".
	pcmSamplingRateIndex	Set to 0x03, indicating a sampling frequency of 48kHz.
	pcmBitsPerSampleIndex	Set to "0", indicating 16-bit PCM data.
	pcmFrameSizeIndex	Indicates a pcmFrameSize equal with the decoder output frame length, and may be set to: "2" if decoder output frame length is set to 1024, "3" if decoder output frame length is set to 2048, "5" for any other decoder output frame length. The decoder output frame length is determined by the core decoder frame length (given by outputFrameLength bit stream element) and the resampling ratio according to Table 10 in ISO/IEC 23008-3 [2].
pcmDataPayload()	bsNumPcmSignalsInFrame	Same value as bsNumPcmSignals.

Additionally, the "PCM to MHAS Embedder" module embeds the MHAS packets of type PACTYP_EARCON, PACTYP_PCMCONFIG and PACTYP_PCMDATA into the received MHAS packet stream on the fly according to the following rules:

- The PCM data is transmitted as a sequence of so called "earcon frames", where each earcon frame, except the last one, consists of the following MHAS packets in the following order:
  - PACTYP_EARCON

  - PACTYP_PCMCONFIG
  - PACTYP_PCMDATA
- The last earcon frame indicates the end of one earcon sequence and consists of the following MHAS packets in the following order:
  - PACTYP_EARCON
  - PACTYP_PCMCONFIG
  - PACTYP_AUDIOTRUNCATION
  - PACTYP_PCMDATA
- The PACTYP_AUDIOTRUNCATION indicating the end of the earcon sequence uses the following:
  - The isActive field in the audioTruncationInfo() structure is set to "1".
  - The truncFromBegin field in the audioTruncationInfo() structure is set to "0".
- The content of PACTYP_EARCON and PACTYP_PCMCONFIG does not change during one sequence of earcon frames.
- The MPEG-H stream created by the "PCM to MHAS Embedder" module can contain a maximum of one earcon frame between two consecutive MHAS packets of type PACTYP_MPEGH3DAFRAME.
- The MHAS packet label for all MHAS packets belonging to the earcon sequence: PACTYP_EARCON, PACTYP_PCMCONFIG, PACTYP_AUDIOTRUNCATION and PACTYP_PCMDATA is "2049".

The MPEG-H Audio decoder part of the external device decodes and renders the PCM data as described in ISO/IEC 23008-3 Amendment 1, Clause 28 [3].

– End of Document –